

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Group Art Unit: *Unassigned*
Examiner: *Unassigned*

In Re PATENT APPLICATION of:

Applicant(s) : Sayori SHIMOHATA, et al

Application No. : *To Be Assigned*

Filed : *Concurrently*

Customer No. 26694

For : DOCUMENT PROCESSING DEVICE
AND DOCUMENT PROCESSING
METHOD


Patent Trademark Office

Attorney Docket : 31759-190464

June 23, 2003

SUBMISSION OF PRIORITY DOCUMENT

Commissioner for Patents
P.O. Box 1450
Alexandria, Virginia 22313-1450

Submitted herewith is a certified copy of Application No. JP2002-182883 filed on June 24, 2002 in Japan, the priority of which is claimed in the present application under the provisions of 35 U.S.C. 119.

Respectfully submitted,

Cathleen M. Kunitz
Norman N. Kunitz
Registration No. 20,586
VENABLE LLP
P.O. Box 34385
Washington, D.C. 20043-9998
Telephone: (202) 962-4800
Telefax: (202) 962-8300

NNK/elw
#465624

日 本 国 特 許 庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出 願 年 月 日

Date of Application:

2002年 6月24日

出 願 番 号

Application Number:

特願2002-182883

[ST.10/C]:

[JP2002-182883]

出 願 人

Applicant(s):

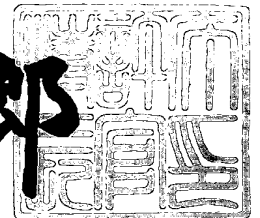
沖電気工業株式会社

Inventor: Sayori SHIMOMURA, et al
ATTN DKT: 31759-190464
Customer NO: 26694

2003年 4月 1日

特 許 庁 長 官
Commissioner,
Japan Patent Office

太田 信一郎



出証番号 出証特2003-3022439

【書類名】 特許願

【整理番号】 KN002523

【提出日】 平成14年 6月24日

【あて先】 特許庁長官 及川 耕造 殿

【国際特許分類】 G06F 15/00

【発明者】

【住所又は居所】 東京都港区虎ノ門1丁目7番12号 沖電気工業株式会
社内

【氏名】 下畑 さより

【発明者】

【住所又は居所】 東京都港区虎ノ門1丁目7番12号 沖電気工業株式会
社内

【氏名】 池野 篤司

【特許出願人】

【識別番号】 000000295

【氏名又は名称】 沖電気工業株式会社

【代表者】 篠塚 勝正

【代理人】

【識別番号】 100090620

【弁理士】

【氏名又は名称】 工藤 宣幸

【手数料の表示】

【予納台帳番号】 013664

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 9006358

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 文書処理装置および方法

【特許請求の範囲】

【請求項 1】 文字情報を含む複数の文書を要素とする集合に関して処理を行う文書処理装置において、

前記集合中の各文書に関し、前記文字情報の共通性を抽出して、前記集合全体に共通の意味内容を表現した文書である共通文書を生成する共通文書生成手段を備えることを特徴とする文書処理装置。

【請求項 2】 請求項 1 の文書処理装置において、

前記共通文書生成手段は、

前記集合中の複数の文書をもとに所定の生成手順を実行して、新たな文書として前記共通文書を生成するか、または、

予め文字情報に共通性のある文書を選んで前記集合を構成した上で、前記集合中の複数の文書のなかから所定の選択手順に応じて 1 つの文書を選択し、選択した当該文書を前記共通文書とすることで前記共通文書を生成することを特徴とする文書処理装置。

【請求項 3】 請求項 2 の文書処理装置において、

前記選択手順では、

前記集合中の各文書に共通して、所定のしきい値以上、出現する頻出表現を検出し、当該頻出表現が最も多く含まれる文書を選択することを特徴とする文書処理装置。

【請求項 4】 請求項 2 の文書処理装置において、

前記共通文書と、前記集合中の各文書との差分となる文字情報である差分文字情報を抽出する差分文字情報抽出手段、または、

前記共通文書と、前記集合中の各文書との共通の文字情報である共通文字情報を抽出する共通文字情報抽出手段を備えることを特徴とする文書処理装置。

【請求項 5】 請求項 4 の文書処理装置において、

前記文書を画面表示する表示手段と、

前記集合の要素である各文書に論理構造を付与し、当該論理構造の付与に際して、少なくとも前記差分文字情報または共通文字情報のいずれかに関しては、その旨を明示する論理構造付与手段と、

当該論理構造付与手段で論理構造を付与した後の各文書を蓄積する文書蓄積手段とを備え、

前記表示手段による画面表示では、当該論理構造を利用した画面表示を行うことを特徴とする文書処理装置。

【請求項 6】 請求項 5 の文書処理装置において、

前記表示手段は、

前記共通文書と当該共通文書に対する各文書の差分文字情報とを、ユーザからの操作に応じて画面表示することを特徴とする文書処理装置。

【請求項 7】 請求項 6 の文書処理装置において、

前記表示手段は、

少なくとも前記差分文字情報を含む文書に関し、オンラインまたはオフラインの出所を示す出典情報を、ユーザからの操作に応じて画面表示することを特徴とする文書処理装置。

【請求項 8】 請求項 7 の文書処理装置において、

前記表示手段は、

ユーザが画面表示された出典情報に対して所定の操作を行うと、前記論理構造を利用して、その出典情報に対応する文書と、前記共通文書との差分文字情報または共通文字情報を識別する所定の識別表示を、画面表示される共通文書に対して実行することを特徴とする文書処理装置。

【請求項 9】 文字情報を含む複数の文書を要素とする集合に関して処理を行う文書処理方法において、

共通文書生成手段が、前記集合中の各文書に関し、前記文字情報の共通性を抽出して、前記集合全体に共通の意味内容を表現した文書である共通文書を生成することを特徴とする文書処理方法。

【請求項 10】 請求項 9 の文書処理方法において、

前記共通文書生成手段は、

前記集合中の複数の文書をもとに所定の生成手順を実行して、新たな文書として前記共通文書を生成するか、または、

予め文字情報に共通性のある文書を選んで前記集合を構成した上で、前記集合中の複数の文書のなかから所定の選択手順に応じて1つの文書を選択し、選択した当該文書を前記共通文書とすることで前記共通文書を生成することを特徴とする文書処理方法。

【請求項11】 請求項10の文書処理方法において、
前記選択手順では、

前記集合中の各文書に共通して、所定のしきい値以上、出現する頻出表現を検出し、当該頻出表現が最も多く含まれる文書を選択することを特徴とする文書処理方法。

【請求項12】 請求項10の文書処理方法において、
前記共通文書と、前記集合中の各文書との差分となる文字情報である差分文字情報を抽出するか、または、

前記共通文書と、前記集合中の各文書との共通の文字情報である共通文字情報を抽出することを特徴とする文書処理方法。

【請求項13】 請求項12の文書処理方法において、
論理構造付与手段が、前記集合の要素である各文書に論理構造を付与し、当該論理構造の付与に際して、少なくとも前記差分文字情報または共通文字情報のいずれかに関しては、その旨を明示し、

文書蓄積手段が、当該論理構造付与手段で論理構造を付与した後の各文書を蓄積し、

表示手段は、当該論理構造を利用した画面表示を行うことを特徴とする文書処理方法。

【請求項14】 請求項13の文書処理方法において、
前記表示手段は、
前記共通文書と当該共通文書に対する各文書の差分文字情報とを、ユーザからの操作に応じて画面表示することを特徴とする文書処理方法。

【請求項15】 請求項14の文書処理方法において、

前記表示手段は、

少なくとも前記差分文字情報を含む文書に関し、オンラインまたはオフラインの出所を示す出典情報を、ユーザからの操作に応じて画面表示することを特徴とする文書処理方法。

【請求項 1 6】 請求項 1 5 の文書処理方法において、

前記表示手段は、

ユーザが画面表示された出典情報に対して所定の操作を行うと、前記論理構造を利用して、その出典情報に対応する文書と、前記共通文書との差分文字情報または共通文字情報を識別する所定の識別表示を、画面表示される共通文書に対して実行することを特徴とする文書処理方法。

【発明の詳細な説明】

【 0 0 0 1 】

【発明の属する技術分野】

本発明は文書処理装置および文書処理方法に関し、例えば、同じキーワードを検索キーとしてテキストデータベースを検索した結果として得られる複数の同類のテキストを処理し、表示する場合などに適用して好適なものである。

【 0 0 0 2 】

【従来の技術】

従来のこの種の装置としては次の文献 1 に開示されるものがある。

文献 1：特開平 9-2 3 1 2 3 8 号公報

文献 1 の表示装置が実行する処理は、テキスト集合を自動的に複数個のグループに分割する分割ステップと、当該分割ステップによって得られた各グループのそれぞれについて、当該グループの属性を表現する主題分類情報を生成する生成ステップと、当該生成ステップで求めた各グループの主題分類情報をグループ別に区分して表示する表示ステップとから構成されている。

ここで、主題分類情報とは、テキストの内容に対応した情報で、キーワードの組あるいは、短い文章を指す。

文献 1 の表示装置ではまた、前記グループと検索条件の間の適合度、および、グループ内の各テキストの、当該グループに対する所属度を算出するステップを有

し、これらの値にしたがって表示するグループやテキストの順番を選択することも可能である。

【0003】

【発明が解決しようとする課題】

しかしながら、上記のような表示装置では、グループごとに提示される各グループの主題分類情報、すなわち、キーワードの組や短い文章からそのグループに含まれるテキストの内容を判断しなければならない。多くの場合、キーワードの組や短い文章だけからそのグループに含まれるテキストの内容（あるいは、グループの概要）を的確に判断することは困難であるため、結局、ユーザはグループに含まれる個々のテキストを読むことによってしか、検索結果を確認することができず、グループの概要を知ることができない。

【0004】

したがって、検索結果を確認したり、グループの概要を知るために長い時間と手数を要し、利便性に欠ける構成となっている。

【0005】

また、上記表示装置において当該主題分類情報が得られるのは、テキスト集合が得られ、さらに当該テキスト集合を分割してグループが得られたあとであるから、テキスト集合が得られた時点では当該主題分類情報さえ存在せず、ユーザがテキスト集合の概要を知るには、個々のテキストを読む以外の方法はなく、極めて利便性が低い。

【0006】

【課題を解決するための手段】

かかる課題を解決するために、第1の発明では、文字情報を含む複数の文書を要素とする集合に関して処理を行う文書処理装置において、前記集合中の各文書に関し、前記文字情報の共通性を抽出して、前記集合全体に共通の意味内容を表現した文書である共通文書を生成する共通文書生成手段を備えることを特徴とする。

【0007】

また、第2の発明では、文字情報を含む複数の文書を要素とする集合に関して

処理を行う文書処理方法において、共通文書生成手段が、前記集合中の各文書に関し、前記文字情報の共通性を抽出して、前記集合全体に共通の意味内容を表現した文書である共通文書を生成することを特徴とする。

【 0 0 0 8 】

【発明の実施の形態】

(A) 実施形態

以下、本発明にかかる文書処理装置および方法を、検索エンジンを含む閲覧システムに適用した場合を例に、実施形態について説明する。

【 0 0 0 9 】

(A-1) 第1の実施形態の構成

本実施形態にかかる閲覧システム10の全体構成例を、図1に示す。図1の各構成要素1～5は、イントラネット内や、あるいは1つの情報処理装置の内部に配置されるものであってもよいが、ここでは、インターネット上に分散配置されるものとして説明する。

【 0 0 1 0 】

図1において、当該閲覧システム10は、入出力部1と、テキストデータベース2と、検索エンジン3と、テキスト集合蓄積部4と、テキスト加工処理部5と、作業用データベース6とを備えている。

【 0 0 1 1 】

このうち入出力部1は、当該閲覧システム10を利用するユーザU1の操作する通信端末に相当する部分で、ハードウェア的には例えばキーボードやマウスなどのポインティングデバイス、ディスプレイ装置、ハードディスクやメモリ装置などを有するパーソナルコンピュータ等が該当し、ソフトウェア的には、当該パーソナルコンピュータが搭載するブラウザ等が該当し得る。

【 0 0 1 2 】

ブラウザとしてはWebページを閲覧するためのWebブラウザがよく知られているが、単にブラウザと呼ぶときは、必ずしもWebブラウザにかぎらず、何らかの情報を閲覧する機能を持つソフトウェア全般を指す。

【 0 0 1 3 】

検索エンジン 3 は、ユーザ U 1 の操作に応じて入出力部 1 から供給される 1 または複数のキーワードをもとに全文検索を実行する部分である。

【 0 0 1 4 】

全文検索とは文書中のすべての文字列から目的の文字列を検索する操作をいう。したがって、例えば、新聞の内容を記述した W e b ページを検索する場合には、当該 W e b ページを構成する H T M L ファイル中の全文字列が検索の対象になる。

【 0 0 1 5 】

全文検索の機能は、必要ならば前記入出力部 1 を有するパーソナルコンピュータに搭載してもかまわないが、W e b (W W W) 上ならば、専門の検索サービス業者がすでに提供している検索サービスを利用することができる。

【 0 0 1 6 】

テキストデータベース 2 は、ハードウェア的には、ハードディスクや光ディスクなどの記憶装置を利用して、多数のテキストを蓄積しているデータベースである。ここで、テキストとは、文書（ドキュメント）と同義である。文書には、データ形式がテキスト形式であるテキストデータ（プレーンテキスト形式のデータ）のほか、G I F や J P E G 等の画像データなども含まれ得る。通常の 1 つの W e b ページは、基本となる 1 つの H T M L ファイル（データ形式として、H T M L 形式はテキスト形式の一種である）のほかに、1 または複数の画像ファイルなどによって構成され得るので、この文書に該当し得る。

【 0 0 1 7 】

この意味で、テキストデータベース 2 は、各種の W e b ページを提供する 1 または複数の W e b サーバと見ることができる。

【 0 0 1 8 】

また、W e b 上の検索サービス業者が検索の対象としているのは、世界中の W e b ページであることからすると、テキストデータベース 2 は、世界中に分散配置された膨大な数の W e b ページ（W e b サーバ）によって構成される W e b （ワールド・ワイド・ウェブ）そのものであると見ることもできる。

【 0 0 1 9 】

もちろん、テキストデータベース 2 は、テキスト（文書）を蓄積するデータベースであるから、Web ページ以外の文書（例えば、XML で記述された文書や、PDF などの電子出版用のデータ形式で記述された文書など）が含まれていてもかまわない。

【 0 0 2 0 】

HTML 形式では、文字の位置や大きさなどを情報の送り手側が詳細に指定することが難しく、色彩の表現力などの点でも、通常の紙媒体の出版物（雑誌や書籍など）に比べるとかなり劣るため、インターネット上の出版物には、送り手側の意思をより忠実に反映することが可能な PDF 形式などが利用されることが多い。なお、PDF 形式で記述された文書は、通常の Web ブラウザの機能だけでは閲覧できないため、入出力部 1 が通常の Web ブラウザだけしか搭載していない場合には、Web ブラウザの機能を拡張するプラグインソフトを搭載することが必要になる。

【 0 0 2 1 】

PDF 形式など、通常のテキスト形式と異なるデータ形式で記述されたファイルは、検索の対象とする前にテキスト形式に変換しておくこと等により、容易に、検索エンジン 3 の検索対象とすることができる。

【 0 0 2 2 】

また、画像データとして文字が記述されることもあり得るが、このような文字も、適宜、テキスト形式に変換することによって検索エンジン 3 による検索の対象とすることが可能である。

【 0 0 2 3 】

テキスト加工処理部 5 は、検索エンジン 3 が前記キーワードを用いた検索の結果として得た複数の文書を加工する部分で、加工後の文書は、テキスト集合蓄積部 4 に蓄積する。本実施形態では、検索エンジン 3 による検索の結果として、内容の類似した複数の文書が得られる場合を想定する。具体的には、例えば、同一の事件に関して記述した同日付けの異なる新聞社による新聞記事などは、ここでいう内容の類似した複数の文書に該当し得る。

【 0 0 2 4 】

一般的には、1つの検索に関し、検索エンジン3に供給するキーワードの数が多いほど、また個々のキーワードが特徴的で識別性が高いものであるほど、検索結果として得られる複数の文書の内容は類似したものとなる傾向がある。検索の結果として得られる文書数は偶発的で予測困難な事象であるから、1つの文書しか得られない可能性もあるが、テキストデータベース2に蓄積されている文書の数十分に多ければ、多くの場合、複数の文書が得られる。

【 0 0 2 5 】

本実施形態では、検索エンジン3による検索の結果として得られた内容の類似した複数の文書は、1つのテキスト集合（文書集合）を構成するものと考え、当該テキスト集合をテキスト加工処理部5の処理の対象とする。なお、当該テキスト集合は、前記文献1の用語との関係では、前記グループではなく、前記テキスト集合に相当する概念である。

【 0 0 2 6 】

（A-1-1）テキスト加工処理部の内部構成例

図1に示すように、当該テキスト加工処理部5は、主題情報生成部5Aと、差分情報生成部5Bと、情報提示部5Cとを備えている。

【 0 0 2 7 】

このうち主題情報生成部5Aは、1つのテキスト集合中の全文書の内容をもとに主題情報を生成する部分である。主題情報とは、当該テキスト集合の主題を示すのに十分な内容を備えた文章である。テキスト集合の主題は、基本的に、1つのテキスト集合中の全文書に共通する内容の文章によって表現される。

【 0 0 2 8 】

例えば、1つのテキスト集合TXG1が3つの文書TX1～TX3から構成されている場合、テキスト集合TXG1の主題情報TH1は、文書TX1～TX3のすべてに共通する内容の文章として表現することができる。

【 0 0 2 9 】

本実施形態における主題情報TH1の表現法には大きく分けて2通りの方法がある。その1つは、文書TX1～TX3の内容をもとに、これらの要約となる新たな文書TXAを生成し当該文書（要約）TXAによって主題情報TH1を表現

する方法（要約生成法）であり、もう1つは、文書TX1～TX3のなかから適切な文書を選択し、選択した文書（例えば、TX3）自体で主題情報TH1を表現する方法（代表選択法）である。

【0030】

要約生成法の実現には、例えば、前記文書TX1～TX3に共通する文節を検出し、検出された各文節を組み合わせることによって前記要約TXAを生成する方法など、様々な方法が使用可能であるが、一例として、次の文献2に記載された方法を用いることもできる。

【0031】

文献2：Columbia Multi-document Summarization: Approach and Evaluation

K.R. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, M.Y. Kan, B. Schiffman, S. Teufel DUC'01

代表選択法の実現にも様々な方法が使用可能であるが、例えば、前記文書TX1～TX3に共通して数多く出現する表現（頻出表現）を抽出し、文書TX1～TX3のうち当該頻出表現が最も多く含まれる文書（例えば、TX3）を代表として選択する方法を用いることができる。

【0032】

差分情報生成部5Bは、前記主題情報TH1に対する各文書（要約生成法を用いた場合にはTX1～TX3、代表選択法を用いた場合には、代表として選択された以外の2つの文書（例えば、TX1、TX2））の差分を抽出する部分である。当該差分（差分情報）は、前記頻出表現を含む単位を各文書に共通する共通情報とし、頻出表現を含まない単位を各文書に固有な固有情報とすれば、当該固有情報として抽出される。ここで、単位とは、節、文、段落などの文法上の単位を指す。

【0033】

差分を抽出したあと、各文書中の差分に該当する単位をマークアップ言語のタグの属性によって、当該単位が差分に該当する旨を指定することができる。

【0034】

例えば、当該マークアップ言語がXML（データ形式として、XML形式はテキスト形式の一種である）の場合には、開始タグと終了タグで当該単位を挟み、開始タグの中に記述する属性によって、当該単位が差分に該当する旨を記述することができる。この場合、必要に応じて、差分情報生成部5Bにおいて、HTMLなどからXMLへのデータ形式の変換を実行することになる。当該単位が差分に該当する旨を示し、なおかつ再利用可能な形で保存するためには、もともとテキストデータベース2上の文書がXML文書でそのようなタグや属性がすでに定義されている場合などを除き、通常は、新たなタグや、新たな属性の定義が必要になり、このような定義が許容されるXML形式を利用する必要があるからである。

【0035】

前記文書TX1～TX3を当該XML形式に変換したあとの文書をXX1～XX3と書く。XML形式の文書XX1は前記TX1に対応し、XML形式の文書XX2は前記TX2に対応し、XML形式の文書XX3は前記TX3に対応する。

【0036】

ただしXML文書は、タグを用いて文書の論理構造を示すだけなので、実際に、各XML文書XX1～XX3の入出力部1における表示方法（ユーザU1が各文書を閲覧する場合の見え方（すなわち、スタイル））を定義するには、スタイルシート言語を用いて具体的な表示方法を定義する必要がある。

【0037】

情報提示部5Cは、前記主題情報生成部5Aで得られた主題情報TH1、差分情報生成部5Bで得られたXML文書XX1～XX3などを、入出力部1のブラウザで表示するのに適した所定の表示形態に加工してユーザU1に提示する部分である。

【0038】

したがって、前記スタイルシート言語を用いた表示方法の定義も、この情報提示部5Cで行うようにするとよい。

【0039】

具体的な表示方法については予め決定し、当該表示方法に対応するスタイルシート言語を、当該情報提示部 5 C に付与しておけば、情報提示部 5 C は、前記主題情報 T H 1 や XML 文書 X X 1 ~ X X 3 などが与えられたとき、自動的に、前記表示形態に加工することができる。

【 0 0 4 0 】

前記テキスト集合蓄積部 4 には、スタイルシート言語によって表示方法まで具体的に定義された XML 文書 X X 1 ~ X X 3 を蓄積しておくための記憶装置である。テキスト集合蓄積部 4 としては、前記入出力部 1 を有する通信端末が搭載したハードディスクなどの一部の記憶領域を利用してもよいが、インターネット上でオンラインストレージサービスを提供する事業者の持つストレージサーバなどを利用することもできる。

【 0 0 4 1 】

いずれにしても、主題情報生成部 5 A、差分情報生成部 5 B、情報提示部 5 C などで行う各処理は、著作物である文書（ここでは、T X 1 ~ T X 3）の改変に相当するものと考えられるため、著作権保護の観点から、これらの処理の成果物である文書 X X 1 ~ X X 3 は、ユーザ U 1 以外のものから閲覧することができないような形式で保存しておくことが望ましい。

【 0 0 4 2 】

前記テキスト加工処理部 5 は、前記入出力部 1 を有する通信端末に搭載するようにしてもよいが、インターネット上にサーバとして配置してもよい。

【 0 0 4 3 】

作業用データベース 6 は、当該テキスト加工処理部 5 内の各構成要素 5 A ~ 5 C が処理を進めるために、前記文書 T X 1 ~ T X 3 などの各データを、整理して蓄積しておくためのデータベースである。最終的に前記文書 X X 1 ~ X X 3 が得られ、テキスト集合蓄積部 4 に蓄積されたあと、当該作業用データベース 6 の蓄積内容は、廃棄することが可能である。

【 0 0 4 4 】

なお、ユーザ U 1 が XML 文書 X X 1 ~ X X 3 を正常に閲覧するためには、入出力部 1 のブラウザは XML 対応のブラウザであることを要する。入出力部 1 が

搭載しているブラウザが通常のWebブラウザなど、XML非対応のブラウザである場合には、プラグインソフトを利用して、XML対応の機能を持たせるようにしてもよい。

【0045】

プラグインソフトは、予め静的に入出力部1に搭載しておくほか、インターネット経由で動的に入出力部1に搭載させることも可能である。

【0046】

以下、上記のような構成を有する本実施形態の動作について、図2のフローチャートを参照しながら説明する。

【0047】

図2のフローチャートは、S1～S6の各ステップから構成されている。

【0048】

(A-2) 実施形態の動作

ユーザU1が入出力部1のブラウザで検索エンジン3にアクセスして所望の検索条件を供給すると(S1)、検索エンジン3は当該検索条件に適合する文書を、前記テキストデータベース2から検索する(S2)。

【0049】

ステップS1を実行する際、入出力部1のブラウザに表示される画面は、例えば、図3に示すものであってよい。

【0050】

図3において、当該画面を構成するウィンドウWD1はユーザU1からの入力を受け入れるための入力用の領域ER1と、基本的にユーザU1への出力を返すために使用される出力用の領域ER3に分けられ、入力用の領域ER1には、フィールドFD1と、ボタンBT1が配置され、出力用の領域ER2には、フィールドFD2と、画面切り替えボタンBT2～BT5が配置されている。

【0051】

このうちフィールドFD1は、ユーザU1からの検索キーの入力を受け入れる検索キー入力部である。ここでは、検索キーとして、日付を含む複数のキーワードの入力を許容するものとするが、必要ならば、文書が作成された日付の範囲（

例えば、2002年6月11日以降に作成された文書のなかから検索) など、各種の検索条件を柔軟かつ詳細に指定できるようにしてもよい。

【0052】

検索キー入力部FD1に入力した検索キーの内容が確定し、ユーザU1が「検索開始」ボタンBT1を操作すると、検索エンジン3に、当該検索キーが供給されて検索が実行される。図3の例では、検索キー入力部FD1に、「Z選手」（野球選手の名前）と、「15日」と、「CCチーム戦」の3つのキーワードを入力している。

【0053】

ここでは、当該3つのキーワードに対応する検索結果として、上述した3つの文書TX1～TX3が得られたものとする。

【0054】

ただし本実施形態の場合、単に検索結果である前記文書TX1～TX3をフィールド（検索結果出力部）FD2に表示するのではなく、前記テキスト加工処理部5による処理の結果を、フィールドFD2に表示するため、フィールドFD2に表示が行われるのは、以降の各ステップS3～S6が実行されたあとである。

【0055】

検索エンジン3による検索の結果として得られた前記3つの文書XT1～XT3は、ステップS3で、作業用データベース6内のテキスト情報格納テーブルTB1に蓄積される。

【0056】

テキスト情報格納テーブルTB1の格納内容は、例えば、図4に示すものであってよい。

【0057】

図4において、当該テキスト情報格納テーブルTB1は、2つの列名（属性）、すなわち、「出典情報」と、「テキスト内容」から構成されている。

【0058】

検索エンジン3の検索によって得られた文書TX1～TX3の数が3であることから、当該テキスト情報格納テーブルTB1の行（タプル）の数も3となって

いる。

【0059】

図示の例では、出典情報として、各文書TX1～TX3の出典である新聞の名称と日付が記述されている。これは人間にとって可読でネットワーク以外の一般社会で使用されるオフラインの出典情報の一例である。必要ならば、このようなオフラインの出典情報に替えて、あるいはオフラインの出典情報とともに、オンラインの出典情報も記述するようにしてもよい。オンラインの出典情報としては、各文書TX1～TX3の前記テキストデータベース2上における存在場所を一意に指定する情報、例えば、URL、FQDN、IPアドレスなどを利用することができる。

【0060】

図4中のテキスト内容から明らかなように、これらの文書TX1～TX3は、米国のP野球リーグで、野球選手Z（外野手）の属するBBチームが、CCチームと対戦した試合における当該Z選手の活躍ぶりを報じる同日付けの新聞記事である。したがって、文書TX1～TX3の内容であるテキスト内容は大部分が同じであるが、例えば、B新聞の記事である文書TX2ではこの試合でZ選手が打率を3割4分9厘に落としたことが記載されているのに、A新聞の記事である文書TX1や、C新聞の記事である文書TX3ではそのことに関する記載は存在しないなど、相違点も存在する。

【0061】

次に、前記主題情報生成部5Aが、当該テキスト情報格納テーブルTB1の格納内容をもとに、前記主題情報TH1を生成する（S4）。

【0062】

ここでは、上述した要約生成法と代表選択法のうち、要約生成法を用いて主題情報TH1を生成したものとする。

【0063】

要約生成法で生成された要約TXAは、少なくとも、テキスト加工処理部5における処理が終了するまでの間、作業用データベース6内に前記テキスト情報格納テーブルTB1とともに格納しておく。もちろん、必要ならば、テキスト情報

格納テーブルTB1のなかに、要約TXAの内容を格納しておくための列名を用意してもよい。

【0064】

このあと、当該要約TXAに対する各文書TX1～TX3の差分情報を抽出する(S5)。ここでは、前記単位として、節を使用しているため、前記XML形式への変換や、変換後のXML文書XX1～XX3のテキスト集合蓄積部4への格納などを行ったあと、ユーザU1からの出力要求に応じて入出力部1上で前記検索結果出力部FD2に表示される内容は、例えば、図7に示すようになる。

【0065】

図7において検索結果出力部FD2内には、最上部に、前記主題情報TH1が表示され、その下に、1行において、オフラインの出典情報OF1である「A新聞5月16日」とA新聞5月16日の記事の当該主題情報TH1に対する差分情報EH1、オフラインの出典情報OF2である「B新聞5月16日」とB新聞5月16日の記事の当該主題情報TH1に対する差分情報EH2、オフラインの出典情報OF3である「C新聞5月16日」とC新聞5月16日の記事の当該主題情報TH1に対する差分情報EH3がそれぞれ表示されている。

【0066】

文書(例えば、XX1)のなかから、差分情報(ここでは、EH1)だけを抽出して、例えば図7に示すように画面表示を行う処理は、前記タグの属性に各単位が差分に該当する旨を指定しておくことによって、入出力部1上のXML対応ブラウザ(あるいは、前記プラグインを装備したブラウザ)の機能だけで実行可能である。

【0067】

文書XX1～XX3中の差分に該当する単位は、図5にアンダーラインを付して示した部分である。

【0068】

図7の画面が入出力部1に表示されると、ユーザU1は個々の文書XX1～XX3の内容を読まなくても、主題情報TH1を読むだけで、テキスト集合の主題を正確に認識することができる。主題情報TH1の文字数は文書XX1～XX3

のうちの任意の1文書の文字数とほぼ同程度であるので、個々の文書XX1～XX3を読む場合に比べ、ユーザU1が読むべき文字数はほぼ1/3程度となる上、各文書XX1～XX3の記事内容の異同をユーザU1の頭脳などを用いて分析する必要もなく、入出力部1上へ個々の文書XX1～XX3のファイルをダウンロードしたり、開いたりするための操作を逐一おこなう必要もない。

【0069】

このためユーザU1は極めて簡単に主題情報TH1を認識することができる。また、これらの効果は、一般的に、1つのテキスト集合中の文書の数が多くなればなるほど、顕著になる。

【0070】

図7の画面例は、ユーザU1が「主題&差分情報表示」ボタンBT4を操作して出力要求を行った場合に対応する表示画面であるが、ユーザU1が「主題&参照情報表示」ボタンBT3を操作して出力要求を行ったときには、図6に示す表示画面が表示される。この参照情報は、前記出典情報に等しい。

【0071】

図6では、前記差分情報EH1～EH3が消失して、オフライン出典情報OF1～OF3だけが、主題情報TH1の下に表示されている。

【0072】

一方、図8は、図7の表示画面上でユーザU1がオフライン出典情報OF3をポインティングデバイスなどを用いて選択したときの表示例を示している。

【0073】

このとき、主題情報TH1上では、随所にアンダーラインが表示され、主題情報TH1の内容のうち当該オフライン出典情報OF3に対応する前記文書TX3から得られた情報がどれであることを直観的に示すことができる。同様に、ユーザU1がオフライン出典情報OF2を選択すれば、アンダーラインが表示されて主題情報TH1の内容のうち当該オフライン出典情報OF2に対応する前記文書TX2から得られた情報を示し、ユーザU1がオフライン出典情報OF1を選択すれば、アンダーラインが表示されて主題情報TH1の内容のうち当該オフライン出典情報OF1に対応する前記文書TX1から得られた情報を示すことができる。

【 0 0 7 4 】

必要に応じて、図 6 の画面上でも、オフライン出典情報を選択することによって同様なアンダーラインを表示するようにしてもよい。

【 0 0 7 5 】

このアンダーラインは、前記スタイルシート言語を変更することによって、反転表示や網かけ表示などへ適宜、変更可能である。また、図 6 ～図 8 における検索結果出力部 F D 2 上のレイアウトなども、スタイルシート言語の変更に応じて変化する。

【 0 0 7 6 】

図 6 ～図 8 のいずれの画面を目視した場合でも、ユーザ U 1 は、主題情報 T H 1 を読むことによって、文書 T X 1 ～ T X 3 （あるいは、 X X 1 ～ X X 3 ）で構成されるテキスト集合の主題を、簡単、かつ確実に認識することが可能である。

【 0 0 7 7 】

必要に応じて、各オフライン出典情報 O F 1 ～ O F 3 と各文書 X X 1 ～ X X 3 （あるいは、テキストデータベース 2 上の各文書 T X 1 ～ T X 3 ）を関連づけておくことにより、オフライン出典情報を選択したときに、当該文書の全文を表示させること等も実行可能である。

【 0 0 7 8 】

（ A - 3 ）実施形態の効果

本実施形態によれば、ユーザ（ U 1 ）は、テキスト集合に含まれる個々の文書（例えば、 T X 1 ～ T X 3 ）を読まなくても、当該テキスト集合の主題（例えば、 T H 1 ）を認識することができ、利便性に優れている。

【 0 0 7 9 】

また本実施形態では、個々の文書と主題との差分情報（例えば、 E H 1 ～ E H 3 ）を表示させたり、主題情報のなかのどの部分（単位）が、各文書に対応しているかを表示させることもできるため、ユーザが各文書を対比したり、分析したりする作業を支援することが可能である。

【 0 0 8 0 】

(B) 他の実施形態

上記実施形態にかかわらず、入出力部 1 の通信端末として、ポインティングデバイス等を備えた一般的なパソコンの代わりにタッチパネル装置を使用したり、専用の通信端末を使用したりすることができる。

【 0 0 8 1 】

また、前記文書 TX 1 ～ TX 3 および XX 1 ～ XX 3 には、単なるテキストデータだけでなく画像データなどが含まれていてもかまわないことはすでに述べた通りである。

【 0 0 8 2 】

なお、上記実施形態では、テキスト加工処理部 5 は、最終的に文書を XML 形式（あるいは、テキスト形式）に変換してテキスト集合蓄積部 4 に蓄積したが、必要に応じて、XML 形式（テキスト形式）以外のデータ形式に変換するようにしてもよいことは当然である。

【 0 0 8 3 】

さらに上記実施形態では、XML のタグや、属性によって、前記単位が差分に該当する旨を示し、なおかつ再利用可能な形で保存するようにしたが、XML のタグや属性以外の方法を用いてこれらの機能を実現してもかまわない。

【 0 0 8 4 】

また、上記実施形態では主題情報 TH 1 の生成にあたり、上述した要約生成法または代表選択法を用いるものとしたが、これら以外の方法で主題情報を生成するようにしてもかまわない。

【 0 0 8 5 】

例えば、テキスト加工処理部 5 が自動的に所定の定型的な手順（例えば、検索された複数の文書（例えば、TX 1 ～ TX 3）のなかから単に文字数の最も少ない文書を主題情報とする）で主題情報を決定するようにしてもよい。

【 0 0 8 6 】

もともと、検索エンジン 3 で検索した時点で文書 TX 1 ～ TX 3 の間の類似度が十分に高い場合などには、このような単純な方法で選択した文書によっても、テキスト集合の主題を、良好に表現することも可能である。

【 0 0 8 7 】

さらに上記実施形態では、主題情報の生成過程にはユーザU1が関与することができず、テキスト加工処理部5側が自動的に生成したが、ユーザU1の意思に応じて主題情報を生成することも可能である。

【 0 0 8 8 】

例えば、前記テキスト集合中の任意の1文書をユーザU1が主題情報として選択できるようにしてもよい。

【 0 0 8 9 】

この場合、ユーザU1の選択に応じて、テキスト加工処理部5が動作し、ユーザU1が選択した1文書と他の文書との差分情報などが自動的に得られる。このような構成は、相互に類似した複数の文書間で、共通点や相違点を詳細に整理する必要がある場合に有用である。

【 0 0 9 0 】

また、上記実施形態にかかわらず検索エンジン3は省略可能である。

【 0 0 9 1 】

現実の文書処理の局面では、検索エンジン3で検索しなくても、予め複数の文書（例えば、TX1～TX3）が与えられているケースも多いからである。また、文書（例えば、TX1～TX3）は必ずしもネットワーク経由で供給されるものである必要はない。例えば、フロッピディスクやCD-ROMなどの記録媒体に格納された形で供給されたり、あるいは、紙媒体の形で供給されたものがOCR処理などを経てシステム内に取り込まれる場合もあってよい。

【 0 0 9 2 】

また、上記実施形態では、同じ試合における野球選手Zの活躍を報じる同日付けの新聞記事であったため、文書TX1～TX3の内容が類似していることが明確に予測できる場合であったが、類似しているか否かが不明な複数の文書に対して本発明を適用してもよい。

【 0 0 9 3 】

その場合、本発明を利用して、文書間の類似度を判定する作業を容易化することが可能になる。

【 0 0 9 4 】

なお、上記実施形態で使用したテキスト情報格納テーブル T B 1 のスキーマは、上述したものに限定する必要はない。テキスト情報格納テーブル T B 1 中の列名を他の列名に置換してもよく、テキスト情報格納テーブル T B 1 中に存在しない列名を追加してもよい。このようなテキスト情報格納テーブルを、必要に応じて、正規化してもよいことは当然である。

【 0 0 9 5 】

さらに、前記作業用データベース 6 とテキスト集合蓄積部 4 は、ハードウェア的には必ずしも別個に設ける必要はなく、統合可能である。

【 0 0 9 6 】

また、上記実施形態にかかわらず、前記入出力部 1 は省略可能である。

【 0 0 9 7 】

例えば、予め与えられたプログラム等にしがって、検索エンジン 3 による検索や、テキスト加工処理部 5 による処理を行い、最終結果である文書（例えば、X X 1 ～ X X 3 ）を、記録媒体に書き込むこと等で処理が完結するシステムもあり得るからである。

【 0 0 9 8 】

また、上記実施形態では、図 3、図 6 ～ 図 8 に具体的な表示画面例を示したが、本発明の表示画面の構成は図示したものに限らないことは当然である。

【 0 0 9 9 】

さらに、前記文書 T X 1 ～ T X 3 は、新聞記事であったが、本発明が対象とする文書が新聞記事にかぎらないことは当然である。

【 0 1 0 0 】

以上の説明では主としてソフトウェア的に本発明を実現したが、本発明はハードウェア的に実現することも可能である。

【 0 1 0 1 】

【発明の効果】

以上に説明したように、本発明の文書処理装置および方法は、従来よりも、利便性に優れている。

【図面の簡単な説明】

【図 1】

実施形態に係る閲覧システムの全体構成例を示す概略図である。

【図 2】

実施形態の動作を示すフローチャートである。

【図 3】

実施形態の動作を示す表示画面例である。

【図 4】

実施形態で使用するテキスト情報格納テーブルの内容例を示す概略図である。

【図 5】

実施形態で使用するテキスト情報格納テーブルの内容例を示す概略図である。

【図 6】

実施形態の動作を示す表示画面例である。

【図 7】

実施形態の動作を示す表示画面例である。

【図 8】

実施形態の動作を示す表示画面例である。

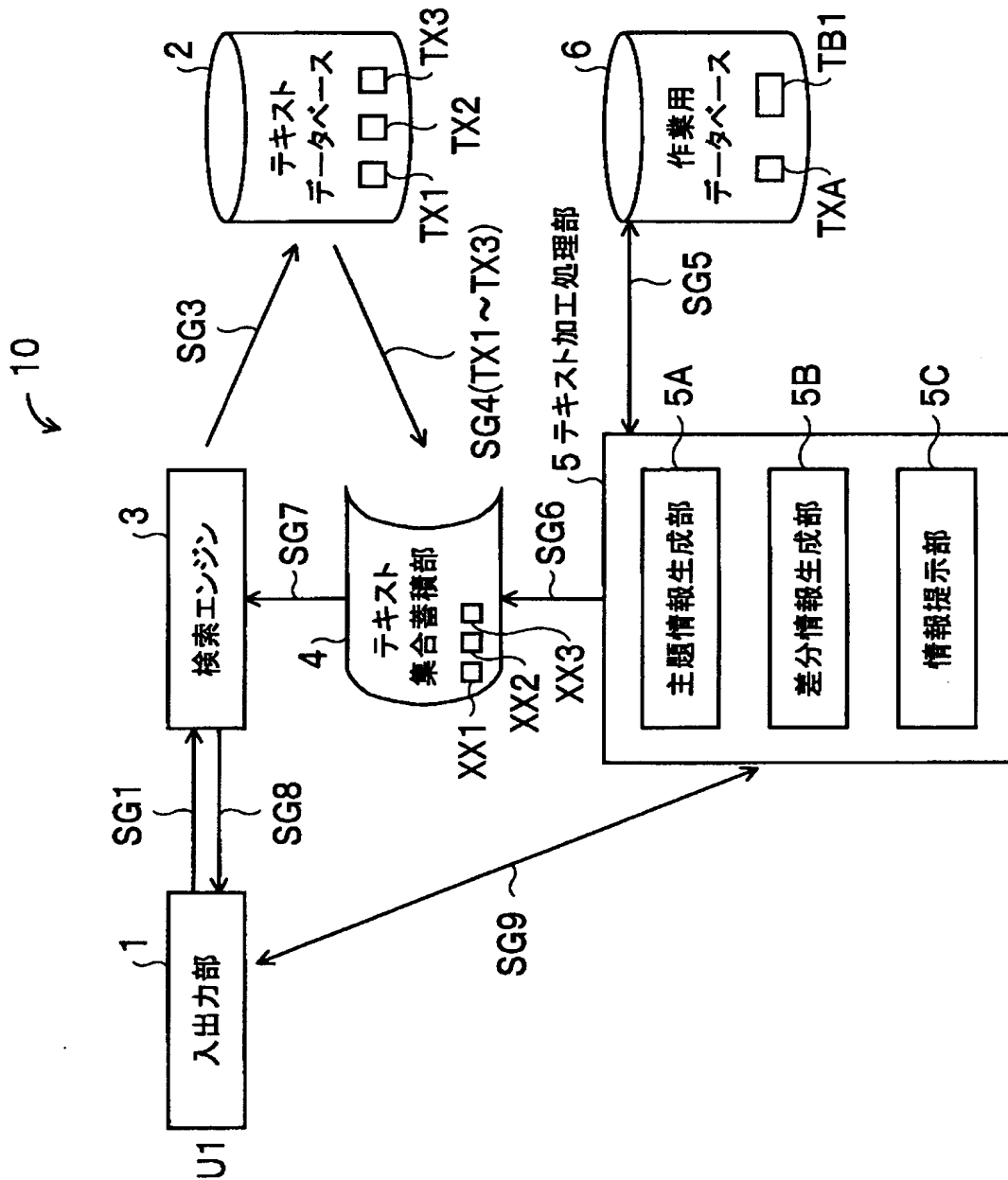
【符号の説明】

1 … 入出力部、2 … テキストデータベース、3 … 検索エンジン、4 … テキスト集合蓄積部、5 … テキスト加工処理部、5 A … 主題情報生成部、5 B … 差分情報生成部、5 C … 情報提示部、6 … 作業用データベース、T B 1 … テキスト情報格納テーブル、T X 1 ～ T X 3 … 文書、X X 1 ～ X X 3 … XML 文書、T X A … 要約、T H 1 … 主題情報（主題）、O F 1 ～ O F 3 … オフライン出典情報、E H 1 ～ E H 3 … 差分情報。

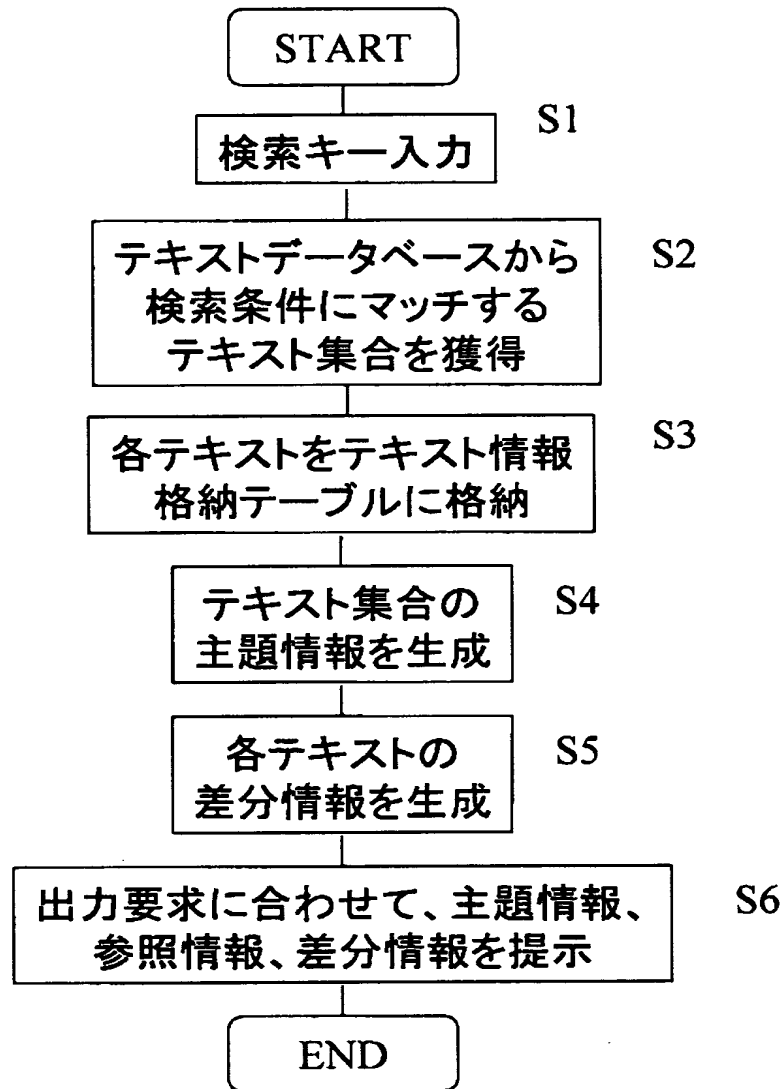
【書類名】

図面

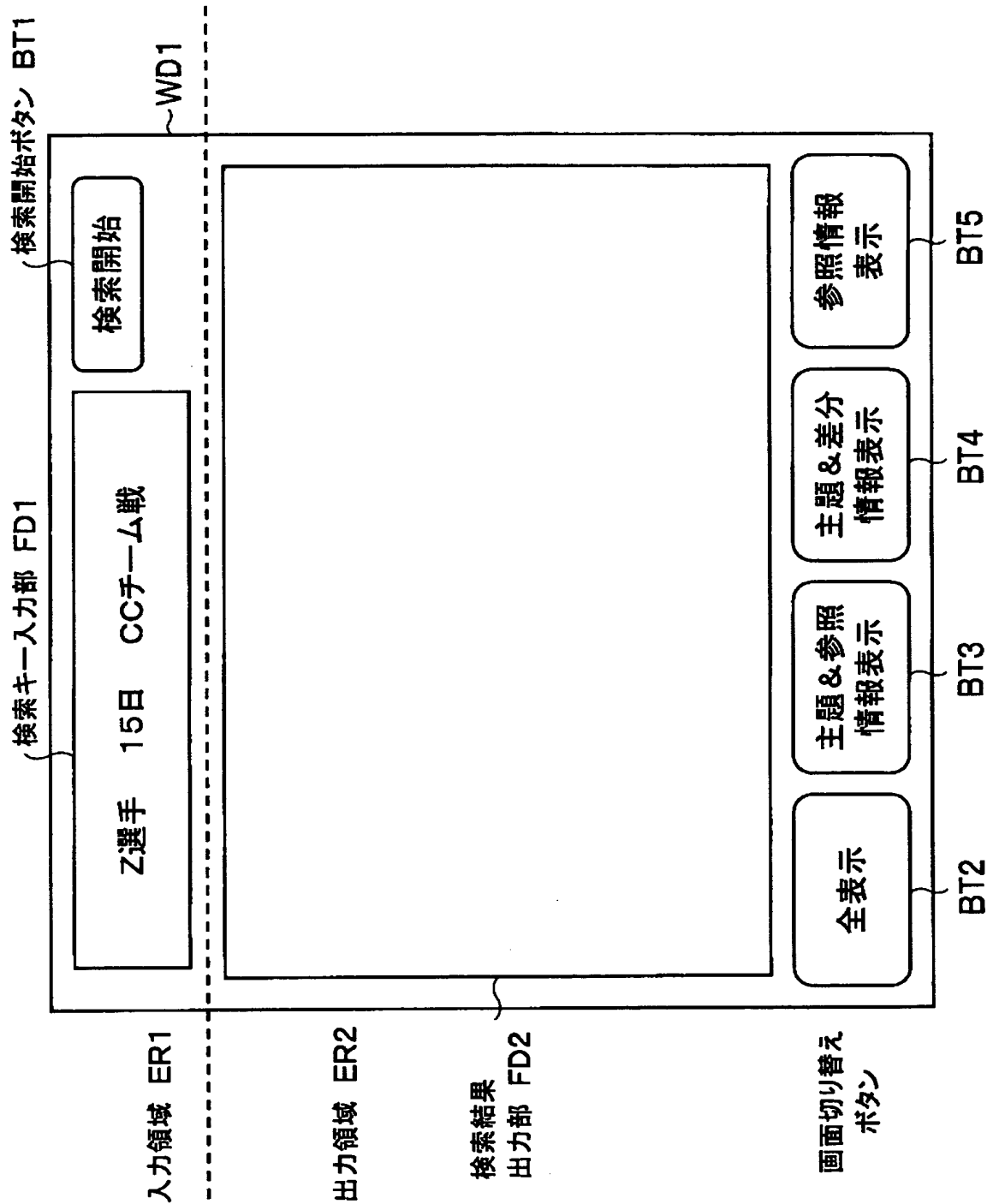
【図 1】



【図 2】



【図3】



【図 4】

テキスト情報格納テーブル TB1

出典情報	テキスト内容
A新聞5月16日	P野球リーグ、BBチームのZ選手は15日(日本時間16日)、カナダ・トロントでのCCチーム戦に1番右翼で先発。4打数1安打1打点で、1盗塁を決めた。Z選手は1回の第1打席は二ゴロ、4回の第2打席は三邪飛だったが、5回の第3打席では、二死一、二塁で左前安打を放った。7回の第4打席は遊ゴロ。9回の第5打席は四球で出塁し、二盗した。試合はBBチームが8-6で勝った。
B新聞5月16日	米P野球リーグ、BBチームのZ外野手は15日、トロントでのCCチーム戦に「1番・右翼」で出場、4打数1安打1打点で、打率を3割4分9厘に落とした。Z選手は二死一、二塁の好機で回ってきた五回の第3打席で、左前適時打を放ち、打点1を挙げた。九回の第5打席では四球を選び、今期12個目の盗塁。Y投手が8-6と逆転した九回裏に登板、無失点で投げきり9セーブ目をあげた。
C新聞5月16日	米P野球リーグ、BBチームのZ外野手は15日、トロントで行われたCCチーム戦で4打数1安打、1打点1得点だった。Z選手は2打席凡退後の五回二死一、二塁の第3打席で左前適時打。6-6の九回の第5打席は四球で出塁し、W選手の決勝適時二塁打につなげた。BBチームのY投手は8-6の九回裏に登板。二死満塁のピンチを招いたが無失点で切り抜け、9セーブ目(2勝)。

TX1

TX2

TX3

【図 5】

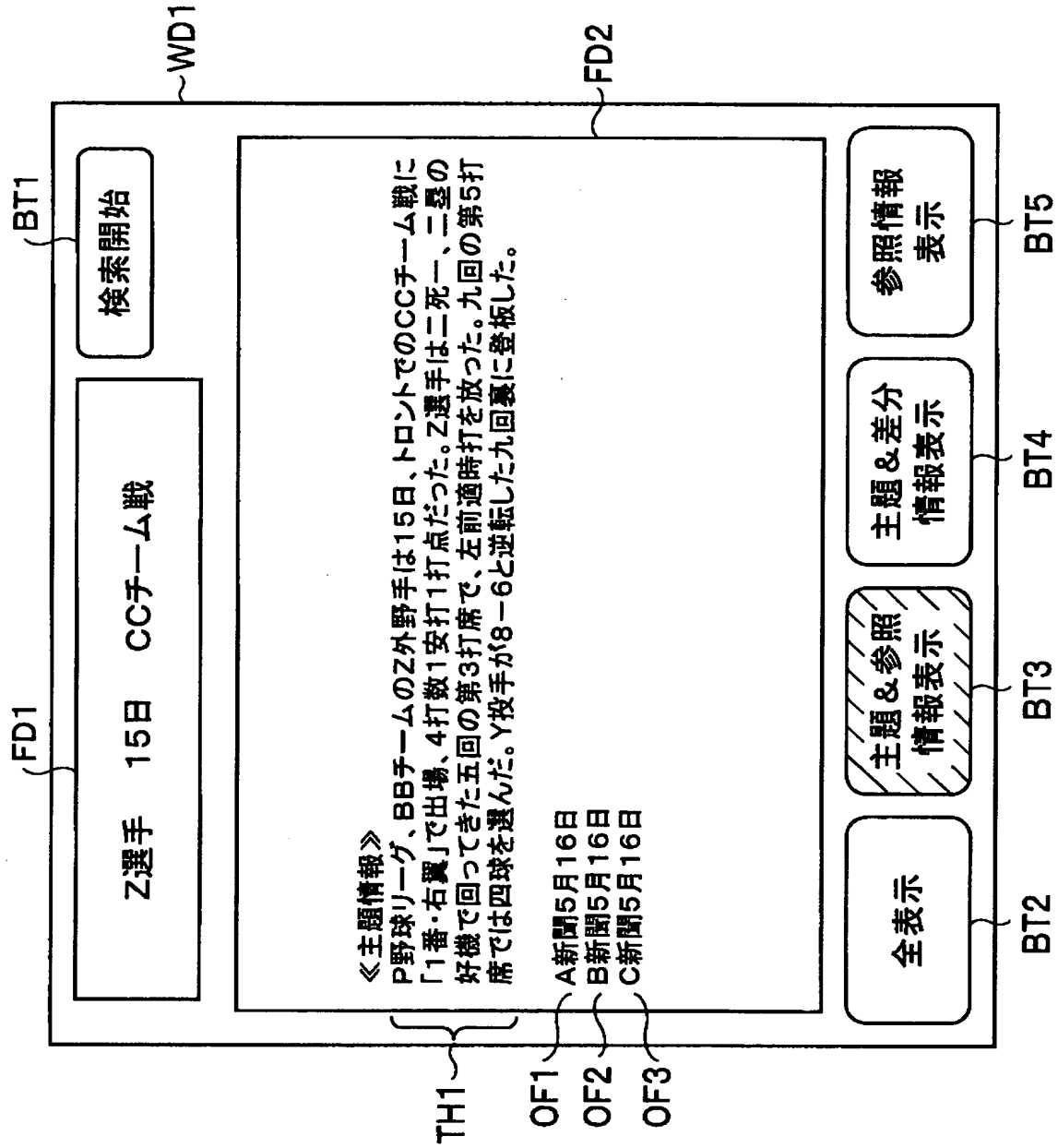
出典情報	テキスト内容
A新聞5月16日	P野球リーグ、BBチームのZ選手は15日(日本時間16日)、カナダ・トロントでのCCチーム戦に1番右翼で先発。4打数1安打1打点で、 <u>1盗塁を決めた</u> 。Z選手は1回の第1打席は二ゴロ、4回の第2打席は三邪飛だったが、5回の第3打席では、二死一、二塁で左前安打を放った。 <u>7回の第4打席は遊ゴロ</u> 。9回の第5打席は四球で出塁し、 <u>二盗した</u> 。試合はBBチームが8-6で勝った。
B新聞5月16日	米P野球リーグ、BBチームのZ外野手は15日、トロントでのCCチーム戦に「1番・右翼」で出場、4打数1安打1打点で、 <u>打率を3割4分9厘に落とした</u> 。Z選手は二死一、二塁の好機で回ってきた五回の第3打席で、 <u>左前適時打を放ち、打点1を挙げた</u> 。九回の第5打席では四球を選び、 <u>今期12個目の盗塁</u> 。Y投手が8-6と逆転した九回裏に登板、 <u>無失点で投げきり9セーブ目をあげた</u> 。
C新聞5月16日	米P野球リーグ、BBチームのZ外野手は15日、トロントで行われたCCチーム戦で4打数1安打、1打点1得点だった。Z選手は2打席凡退後の五回二死一、二塁の第3打席で左前適時打。6-6の九回の第5打席は四球で出塁し、 <u>W選手の決勝適時二塁打につなげた</u> 。BBチームのY投手は8-6の九回裏に登板。 <u>二死満塁のピンチを招いたが無失点で切り抜け、9セーブ目(2勝)</u> 。

XX1

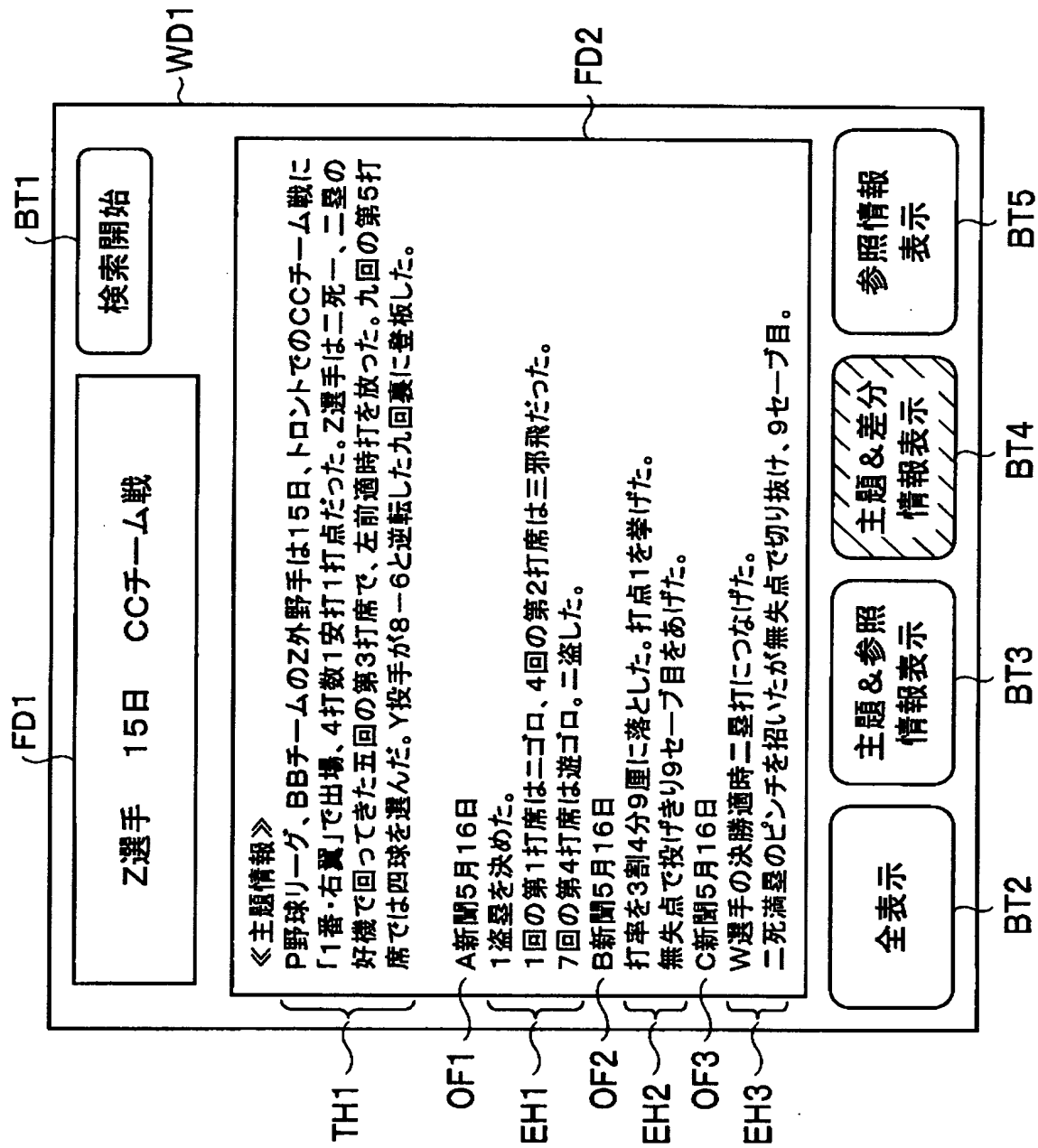
XX2

XX3

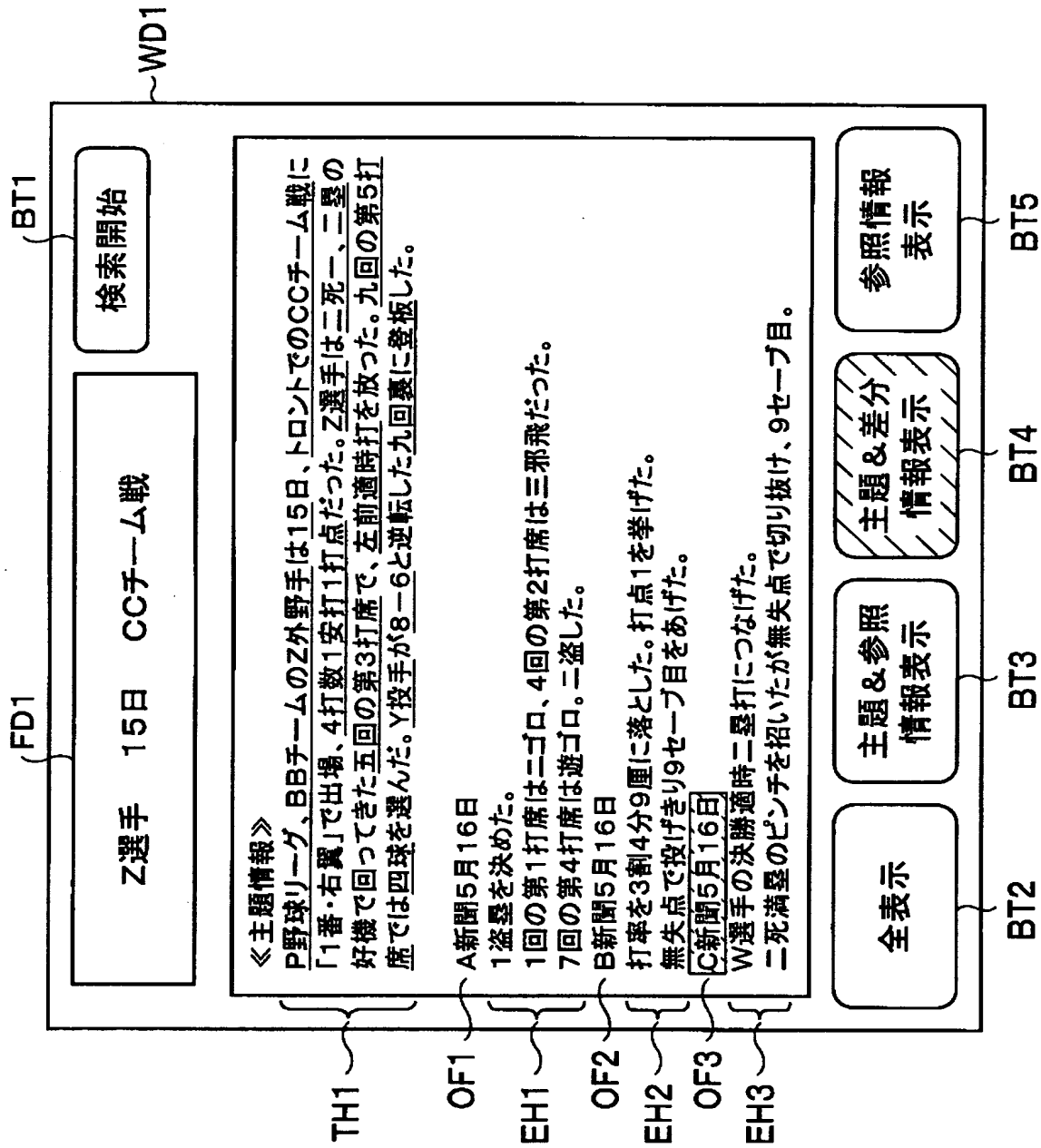
【図6】



【図7】



【図 8】



【書類名】 要約書

【要約】

【課題】 利便性を高める。

【解決手段】 文字情報を含む複数の文書を要素とする集合に関して処理を行う文書処理装置において、前記集合中の各文書に関し、前記文字情報の共通性を抽出して、前記集合全体に共通の意味内容を表現した文書である共通文書を生成する共通文書生成手段を備える。またこの文書処理装置において、前記共通文書生成手段が、前記集合中の複数の文書をもとに所定の生成手順を実行して、新たな文書として前記共通文書を生成するか、または、予め文字情報に共通性のある文書を選んで前記集合を構成した上で、前記集合中の複数の文書のなかから所定の選択手順に応じて1つの文書を選択し、選択した当該文書を前記共通文書とすることで前記共通文書を生成することは好ましい。

【選択図】 図1

出 願 人 履 歴 情 報

識別番号 [000000295]

1. 変更年月日 1990年 8月22日
[変更理由] 新規登録
住 所 東京都港区虎ノ門1丁目7番12号
氏 名 沖電気工業株式会社